基于语义分布相似度的主题模型

居亚亚，杨　璐，严建峰

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

摘　要：潜在狄利克雷分布(LDA)是一种流行的三层贝叶斯概率模型，其实现了文本与文本中的单词在主题层次上的聚类。LDA 以词袋(Bag of Words，BOW)模型为基础，简化了建模的复杂度，但使得主题的语义连贯性较差，文档表征能力不强。为解决此问题，提出了一种基于语义分布相似度的主题模型。该模型在 EM(Expectation Maximization)算法框架下，使用 GPU(generalized Pólya urn)模型加入单词-单词和文档-主题语义分布相似度来引导主题建模，从语义关联层面上削弱了词袋假设对主题产生的影响。在四个公开数据集上的实验表明，基于语义分布相似度的主题模型在主题语义连贯性、文本分类准确率方面相对于目前流行主题建模算法表现的更加优越，同时该模型提高了收敛速度和模型精度。

关键词：潜在狄利克雷分布；语义分布相似度；主题模型；GPU 模型

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.07.0385

Semantic Distribution Similarity Based Topic Model

Ju Yaya, Yang Lu, Yan Jianfeng

(*School of Computer Science &Technology*, *Soochow University*, *Suzhou Jiangsu* 215006, *China*)

Abstract: The latent Dirichlet allocation (LDA) is a popular three-layer Bayesian probability model that implements clustering of words in text and text at the topic level. LDA is based on the bag-of-words, which simplifies the complexity of modeling, but makes the semantic coherence of topics poor, and text representation ability is not strong. To solve this problem, this paper came up with the semantic distribution similarity based topic model. This model uses GPU (generalized Pólya urn) model to add word-word and document-topic semantic distribution similarity to guide topic modeling under the framework of EM (Expectation Maximization) algorithm, which weakened the effect of bag-of-words hypothesis on topics from the semantic association level. Experiments on four public datasets show that the semantic distribution similarity based topic model is superior to the currently popular topic modeling algorithms in terms of topic semantic coherence and text classification accuracy, and the model improves the convergence speed and topic accuracy.

Key words: latent Dirichlet allocation; semantic distribution similarity; topic model; GPU model

0　引言

当前，随着互联网技术的高速发展，网络数据呈现爆炸式的增长，主要包括微博、新闻、网页、图像和声音等，其中网络中的文本信息占据着主要的地位，如何从海量文本信息中获取所需要的知识是人们目前所面临的一大难题，其中主题模型是解决这一难题的有效工具，主题模型是一种利用非监督的机器学习算法来抽取隐藏在文档和单词中的潜在主题信息的统计模型，其中潜在狄利克雷分布(Latent Dirichlet Allocation，LDA)[1]是一种常用的概率主题模型，通过将主题作为文档与单词之间的中间层特征的表达方式，实现显示地抽取文本的语义信息，常被用于文本分类[1,2]、摘要抽取[3]、主题检测和追踪[4]等任务。

目前 LDA 主题模型流行的推理算法主要有变分贝叶斯(Variational Bayesian，VB)[1]、吉布斯采样(Gibbs Sampling，GS)[5]和期望最大化(Expectation Maximization，EM)[6,7]。基于这三种推理算法产生了一些针对特定应用场景的变种算法，如随机变分贝叶斯[8]、作者主题模型[9]、自适应的期望最大化[7]等，虽然这些算法能够取得一定的建模效果，但是在建模过程中仍然存在一系列的挑战。首先，当前的主题模型变种通常是加入外部的先验知识引导建模来实现功能或语义上的增强，如 Chen 等人[10]提出了 GK-LDA 模型(General Knowledge LDA)，通过利用领域独立的通用知识来获取单词间语义关系，并融合到主题建

收稿日期：2018-07-23；**修回日期：**2018-09-13　**基金项目：**国家自然科学基金资助项目（61572339，61272449）；江苏省科技支撑计划重点项目（BE2014005）；

作者简介：居亚亚（1989-），女，江苏徐州人，硕士研究生，主要研究方向为机器学习、数据挖掘（yayaju@163.com）；杨璐（1982-），女，副教授，硕导，主要研究方向为机器学习、软件工程；严建峰（1978-），男，副教授，硕导，主要研究方向为机器学习、数据挖掘.

模过程，提高主题一致性，其主要针对特定领域的短文本任务进行的改进，并不具有普遍性，同时获取的先验知识可能存在错误。其次，当前的主题建模没有较好地结合相关的语义强化机制，如 Bekoulis 等人[11]提出了一种基于图模型的加权方法，其假定文档中两个单词的共现次数越高，相应的权重越大，这样使得主题模型能够从长文档中获得更具区分能力的主题，但是这种加权方法并未在建模中考虑单词之间的语义关系，因此不能获得语义连贯性和可解释性较优的主题[12]。此外，当前主题建模大部分使用吉布斯采样（GS）来实现参数的估计，此算法往往使得模型的迭代不能收敛到一个理想状态，这样使得文档的语义表征能力不强。针对以上问题，本文基于语义分布相似度构建文本主题模型，旨在增强主题的语义连贯性、提高文本分类的准确率，同时提高收敛速度和精度。

本文研究了概率主题模型的语义强化问题，提出了基于语义分布相似度的主题模型(Semantic Distribution Similarity based Topic Model，SDS_TM)，在 EM 算法框架下，使用广义波利亚坛子模型(Generalized Pólya Urn，GPU)[13]从单词-单词和文档-主题这两个方面进行语义强化，并实现 SDS_TM 的参数估计。首先，针对单词-单词的语义强化，通过单词的语义分布表示获得单词之间的相似性；其次，针对文档-主题的语义强化，通过计算文档的语义分布表示和文档中单词的语义分布表示之间的相似性来获得文档语义的代表词，以其数量上的增加来提高文档中相应主题的概率。本文将基于语义分布相似度的主题模型与目前流行的推理算法：变分贝叶斯(VB)、吉布斯采样(Gibbs Sampling)和期望最大化(EM)进行了对比，实验表明，基于语义分布相似度的主题模型能够在主题语义连贯性、文本分类准确性方面表现的更加优越，同时能够有效地提高收敛速度和精度。

1　相关工作

1.1　LDA 主题模型

LDA 模型是一种无监督的三层贝叶斯概率图模型，包含文档、主题、单词三层。LDA 图模型如图 1 所示，其中非阴影圆圈表示参数或需要估计的隐藏变量；阴影的圆圈表示可观测到的变量；箭头表示两变量之间的依赖关系；方框表示重复过程；方框中的下标表示重复的次数。LDA 模型假定整个文本集有 K 个主题，每篇文档 d 可以表示为长度为 K 的主题分布 θ_d，每个主题 k 表示为长度为词汇表长度 W 的单词分布 ϕ_k，一篇文档是生成过程如下：

$$\theta_d \sim Dir(\alpha), \phi_k \sim Dir(\beta), z_i \sim \theta_d, x_i \sim \phi_{z_i} \tag{1}$$

其中假设 θ_d 和 ϕ_k 服从狄利克雷分布(Dir)，其超参数分别为 α 和 β。LDA 的建模过程是逆向的通过文本集合生成模型，首先从先验参数为 β 的狄利克雷分布中获取每个主题 k 的分布 ϕ_k，对于一篇文档 d，从先验参数为 α 的狄利克雷分布中获取其主题分布的概率分布 θ_d，接下来从 θ_d 中采样出文档 d 中每个单词 t 的主题 z_i，再从主题单词分布 ϕ_{z_i} 中获取 w。重复这样的过程直到生成所有的文档。表 1 列出了本文所使用的一些参数。

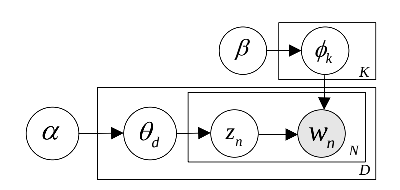


图 1　LDA 图模型

表 1　符号定义

符号	意义
$1 \le d \le D$	语料库文本索引
$1 \le w \le W$	词汇表中单词索引
$1 \le k \le K$	主题索引
$1 \le t \le T$	迭代次数
$x_{w,d}$	索引为 $\{w,d\}$ 的单词词频
x	所有 $x_{w,d}$ 的集合
NNZ	非零元素个数
$z^k_{w,d}$	文本 d 中所有单词 w 属于主题 k 的个数
z	所有 $z^k_{w,d}$ 的集合
θ_d	文本 d 的主题分布
$\theta_d(k)$	文本 d 的主题分布中主题 k 的概率
ϕ_k	主题 k 的单词分布
$\phi_w(k)$	主题 k 的单词分布中单词 w 的概率
$\theta_d(k)$	文本 d 的主题分布中主题 k 的概率计数
$\phi_w(k)$	主题 k 的单词分布中单词 w 的概率计数
$\mu_{w,d}(k)$	文本 d 中单词 w 属于主题 k 的概率
α,β	狄利克雷分布的超参数

LDA 的推理目标是从联合概率分布 $p(\mathrm{x},\mathrm{z},\theta,\phi \mid \alpha,\beta)$ 中最大化特定的后验概率，不同的 LDA 算法对于后验概率的理解不同，目前主流的推理算法主要有变分贝叶斯(VB)、吉布斯采样(GS)和期望最大化(EM)，由于这些推理算法优化不同的后验概率下界，所以其建模的结果存在差异。

1.2　基于变分推断的算法

Blei 在提出 LDA 模型时给出了一种基于变分推断的求参方法(VB)，该算法的核心是利用变分推断方法将无法求解的后验概率分布用可解的近似分布代替，通过近似分布来求解变分参数，通过不断地迭代求出模型参数。其定义的优化目标为：

$$p(\theta,\mathrm{z} \mid \mathrm{x},\phi,\alpha,\beta) \propto p(\theta,\mathrm{z},\mathrm{x},\phi \mid \alpha,\beta) \tag{2}$$

变分推断利用平均场近似(Mean Field Approximation)理论，将近似分布赋予可完全分解的性质。该近似分布定义为：

$$q(\theta,\mathrm{z} \mid \gamma,\delta) = q(\theta \mid \gamma)\prod_{n=1}^{N} q(\mathrm{z}_n \mid \delta_n) \tag{3}$$

其中 γ 和 δ 为文档级的自由参数，简化的 LDA 概率图模型如图 2 所示。

通过最小化近似分布和真实分布之间的 Kullback-Leibler(KL)距离来求导参数值，可得到近似分布的更新公式为：

$$\mu_{w,d}(k) \propto \frac{\exp[\Psi(\theta_d(k)+\alpha)]\exp[\Psi(\phi_w(k)+\beta)]}{\exp[\Psi(\sum_w[\phi_w(k)+\beta])]} \tag{4}$$

其中：

$$\theta_d(k)=\sum_w x_{w,d}\mu_{w,d}(k) \tag{5}$$

$$\phi_w(k)=\sum_d x_{w,d}\mu_{w,d}(k) \tag{6}$$

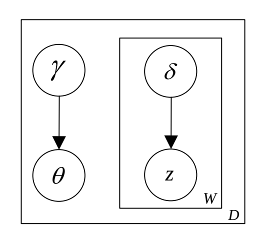


图 2　简化的 LDA 图模型

1.3　基于吉布斯采样的算法

吉布斯采样(GS)是 LDA 模型解决近似推理问题的一种解法，对难以求解的隐变量的联合后验概率进行近似采样。GS 的优化目标为：

$$p(\mathrm{z}\,|\,\mathrm{x},\alpha,\beta)=\frac{p(\mathrm{x},\mathrm{z}\,|\,\alpha,\beta)}{p(\mathrm{x}\,|\,\alpha,\beta)}\propto p(\mathrm{x},\mathrm{z}\,|\,\alpha,\beta) \tag{7}$$

GS 是马尔科夫蒙特卡洛(Markov Chain Monte Carlo，MCMC)[14]算法的一个特例，使用 MCMC 从目标分布中采样，首先，移除当前单词 i 的主题标签 $z_{w,d,i}^{k,old}$ ，然后，根据除单词 i 之外的所有单词的主题标签分布 $z_{w,d,-i}^{k,old}$ 来估计当前单词分配给各个主题的概率 $\mu_{w,d,i}(k)$，最后，随机采样出一个主题标签分配 $z_{w,d,i}^{k,new}=1$ 给当前单词，不断迭代直到收敛。更新公式为：

$$\mu_{w,d,i}(k)\propto\frac{[\theta_d^{-z_{w,d,i}^{k,old}}(k)+\alpha][\phi_w^{-z_{w,d,i}^{k,old}}(k)+\beta]}{\sum_w[\phi_w^{-z_{w,d,i}^{k,old}}(k)+\beta]} \tag{8}$$

从 $\mu_{w,d,i}(k)$ 中采样 $z_{w,d,i}^{k,new}=1$

其中：

$$\theta_d(k)=\theta_d^{-z_{w,d,i}^{k,old}}(k)+\mathrm{z}_{w,d,i}^{k,new} \tag{9}$$

$$\phi_w(k)=\phi_w^{-z_{w,d,i}^{k,old}}(k)+\mathrm{z}_{w,d,i}^{k,new} \tag{10}$$

1.4　基于期望最大化的算法

期望最大化算法类似于 LDA 后验概率最大算法(Maximum A Posterior，MAP)[15]，其优化目标为：

$$p(\theta,\phi\,|\,x,\alpha,\beta)=\frac{p(\mathrm{x},\theta,\phi\,|\,\alpha,\beta)}{p(\mathrm{x}\,|\,\alpha,\beta)}\propto p(\mathrm{x},\theta,\phi\,|\,\alpha,\beta) \tag{11}$$

最大化该后验概率可以理解为寻找拟合 x 的最优 $\{\theta,\phi\}$，将似然概率 $p(\mathrm{x},\theta,\phi\,|\,\alpha,\beta)$ 展开并利用 Jensen 不等式进行最大化，求导后可得到 EM 算法的 EM 框架，其中 E-step 为更新文档 d 中的单词 w 属于主题 k 的概率：

$$\mu_{w,d}(k)\propto\frac{[\theta_d(k)+\alpha-1][\phi_w(k)+\beta-1]}{\sum_w[\phi_w(k)+\beta-1]} \tag{12}$$

M-step 为更新充分统计量 $\{\theta_d(k),\phi_w(k)\}$：

$$\theta_d(k)=\sum_w x_{w,d}\mu_{w,d}(k) \tag{13}$$

$$\phi_w(k)=\sum_d x_{w,d}\mu_{w,d}(k) \tag{14}$$

1.5　目前算法分析对比

以上是目前 LDA 主流的三种推理算法，由于这些算法优化目标是不同隐变量之间的组合，并实现间接地求解 LDA 的参数 $\{\theta_d(k),\phi_w(k)\}$，所以它们之间存在着许多不同点。变分贝叶斯(VB)和吉布斯采样(GS)均使用近似推断的方法实现主题建模，此外，VB 算法在计算主题分布时引入了 digamma 函数，因此，算法的精度较低、收敛速度较慢，然而期望最大化(EM)在求解参数 $\{\theta_d(k),\phi_w(k)\}$ 时使用确切的推断得到后验概率的确切下界，因此该算法在收敛速度和精度上均优于 VB 和 GS 算法[7]，然而，这三种推理算法都是以词袋(BOW)模型为假设，既不考虑文档与文档中单词的关系，也不考虑单词与单词之间的关系，这种假设虽然简化了建模的复杂度，但是使得主题建模的效果不理想。

2　基于语义分布相似度的主题建模

目前流行的 LDA 模型推理算法均是以词袋(BOW)模型为假设，即将文档表示成一个词频向量，这样在建模过程中忽略了文档与单词、单词与单词之间的语义关联，丢失了文档的句法、语法等信息，因此许多研究对主题模型进行了一些扩展，但是这些扩展主要是针对特定任务或者是引入外部先验知识引导主题的建模[16]，都是对传统主题模型应用的扩展或改进，并没有实质性的差别。

本文提出了一种基于语义分布相似度的主题模型，此模型在 EM 算法框架下分别从单词-单词和文档-主题两个方面进行语义强化，主要思想是考虑单词与单词之间的语义关联，即与被采样单词语义关联较强的单词属于相同主题的概率较大，同时还考虑了文档和文档中的单词之间的语义关联，与文档语义关系紧密的单词被该文档相应主题选择的概率增大，即实现了文档-主题的语义强化。通过双向的语义强化对主题建模的过程进行改进，有效地增强了主题语义连贯性和文档表征能力。

2.1　基于 GPU 的语义强化

广义波利亚坛子(generalized Pólya urn，GPU)模型常被用于主题模型的采样过程中，在上下文主题模型中，一个单词被看做一种颜色的球，一个主题被看做一个坛子，主题的分布通过坛子中不同颜色的球的个数来反映，LDA 模型遵循广义波利亚坛子模型的原因是当从坛子中取出特定颜色的球时，则将球与球颜色相同的球一起放回坛子中，随着时间推移，坛子中球的个数变化是一种自我强化的现象，即“富人越来越富”，这个过程与主题模型中单词的主题采样是一致的。本文采用 GPU 模型分别从单词-单词和文档-主题这两个方面进行主题建模的语义强化。

2.1.1 单词-单词的语义强化

以往的大多数研究主要是通过外部先验知识获得单词与单词之间的语义关系，这样获得的语义知识不一定符合建模的语料库，因此，本文在不引入外部先验知识的条件下，从单词的局部上下文语法信息和全局文档范围内的语义信息这两个角度考虑，获得语料库中单词之间的语义关联。

单词-单词的 GPU 语义强化，是通过计算单词语义分布表示之间的余弦相似度来实现。单词的局部语义分布是通过 word2vec[17]模型获取，word2vec 模型将单词表示为一种分布式词向量形式，仅从单词所在位置周围的文档信息考察单词的语义，忽略单词在全局文档中的主题信息，通过固定大小的滑动窗口对语料库中每个单词进行上下文统计，获得单词 w 的上下文语义分布表示 $\boldsymbol{v}_w=\{v_1,v_2,...,v_k\}$，其维度为 K，$v_i\in[0,1]$。单词的全局语义分布表示是通过 LDA 模型产生的主题单词分布 $\phi_w(k)$ 获得，$\phi_w(k)$ 是在语料库上建模产生的全局文档范围内的语义信息，单词 w 的主题分布被表示为一个 K 维向量 $\boldsymbol{\kappa}_w=[\phi_w(1),\phi_w(2),...,\phi_w(K)]$，其中 $\phi_w(k)\in[0,1]$，文献[18]中对词的主题分布向量进行了研究，$\phi_w(k)$ 是一个稀疏矩阵，当 K 足够大时 $|\kappa|=\sqrt{\sum_{k=1}^{K}|\phi_w(k)|}\rightarrow 0$，并且由于词袋(BOW)模型的影响，文档中的高频词具有较低的稀疏性，关键词或低频词具有较高的稀疏性，传统主题建模过程中高频词几乎占据所有的主题，因此，本文在单词语义分布表示中引入 L2 范数来抑制高频词对建模的影响，L2 范数是用来衡量向量的稀疏度，公式(15)是单词的主题向量稀疏度的计算公式，其中 K 表示主题数。

$$\delta_w=\frac{\sqrt{K}-(\sum_k|\phi_w(k)|)/\sqrt{\sum_k[\phi_w(k)]^2}}{\sqrt{K}-1} \tag{15}$$

因此，将单词的局部语义分布 $\boldsymbol{v}_w$ 和全局语义分布 $\boldsymbol{\kappa}_w$ 进行线性加权求和，可得单词 w 的语义分布表示为 $\boldsymbol{t}_w=\boldsymbol{v}_w+\delta_w\boldsymbol{\kappa}_w$，其中权重 δ_w 表示在向量空间中对单词的位置进行了调整，使得同一主题下的单词在向量空间中的距离更近。对于被采样的单词 w，与其余弦相似度大于阈值 $\lambda(0\le\lambda\le 1)$ 的单词构成该单词的相似单词集合 $\boldsymbol{W}_w$，假设单词的相似矩阵为 $\boldsymbol{A}$，当单词 w 被采样时，则集合 $\boldsymbol{W}_w$ 中的所有单词在采样主题上的概率值都将被增加相应的余弦相似度，对于当前单词 w 自身增强不变，仍为 1，其他情况下单词不进行强化。具体的强化方式如式(16)所示。

$$A_{w,w^*}=\begin{cases}1, & w=w^*\\ cos(t_w,t_{w^*}), & w^*\in\boldsymbol{W}_w\text{且}w\neq w^*\\ 0, & \text{其他}\end{cases} \tag{16}$$

2.1.2 文档-主题的语义强化

以往对于主题模型强化的大部分研究仅仅停留在词义相近的单词之间的语义关联，并未考虑单词与文本之间的语义关联。本文从文档的语义分布表示出发，考虑建模产生的文档主题分布与文档中单词的责任值 $\mu_{w,d}(k)$ 之间的语义关联来获得文档语义的代表单词，其中 $\mu_{w,d}(k)$ 是一个稀疏矩阵，所以使用其 L2 范数来约束词袋(BOW)模型对语义强化的影响。单词与所处文档之间的语义关联在 GPU 模型强化过程中的体现是当单词 w 被主题 k 采样时，若该词与文档 d 的语义关联密切，则文档 d 中主题 k 的概率值将被增强。文档-主题的 GPU 语义强化是通过计算语料库中单词语义分布与其所在文档的语义分布之间的语义相似度，如式(17)所示，其中 δ_w 表示单词 w 的稀疏度，$\mu_{w,d}(k)$ 为文档 d 中单词 w 属于主题 k 的概率值，$\theta_d(k)$ 为文档主题分布，d_W 为文档 d 中所有单词的集合。若两者之间的相似度大于阈值 $\rho(0\le\rho\le 1)$，则认为文档 d 与主题 k 之间需要语义的增强 $Sim_{w,d}$，否则不需要进行强化。强化矩阵为 $\boldsymbol{B}$，具体强化的方式如式(18)所示：

$$Sim_{w,d}=\frac{\sum_k\theta_d(k)*(\delta_w\mu_{w,d}(k))}{\sum_{w^{'}\in d_W}[\sum_k\theta_d(k)*(\delta_{w^{'}}\mu_{w^{'},d}(k))]} \tag{17}$$

$$B_{d,k}=\begin{cases}Sim_{w,d}, & Sim_{w,d}>\rho\\ 0, & Sim_{w,d}\le\rho\end{cases} \tag{18}$$

2.2 SDS_TM 模型结构

本文提出了基于语义分布相似度的主题模型(SDS_TM)。SDS_TM 是在 LDA 模型的基础上，采用 GPU 模型融合单词-单词和文档-主题的语义分布相似度来实现主题建模过程中的语义强化。

SDS_TM 的图模型如图 3 所示，图中斜线阴影部分表示文档-主题部分和单词-单词部分的 GPU 语义强化。前者依赖于主题建模中产生的文档主题分布和主题单词分布，后者不仅依赖于主题单词分布，还依赖于 Skip-Gram 词嵌入，即使用 word2vec 中的 Skip-Gram 模型获得的单词局部语义分布。

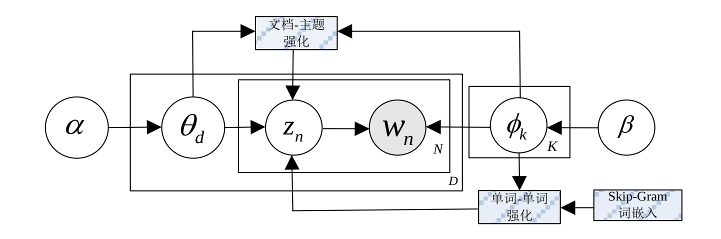


图 3　SDS_TM 图模型

2.3 模型参数推断

目前主流的主题模型推理算法有变分贝叶斯(VB)、吉布斯采样(GS)和期望最大化(EM)算法，其中 EM 算法直接优化的是 LDA 模型后验概率的确切下界，在泛化性能和精度方面较 VB 和 GS 算法表现的更加优越，所以基于 EM 算法框架对 SDS_TM 模型的参数进行推断。根据 EM 算法的更新公式，文档 d 中单词 w 在主题 k 上的更新公式 $\mu_{w,d}(k)$ 如式(19)所示，使用 GPU 模型融合单词-单词和文档-主题的语义分布相似度，充分统计量 $\theta_d(k)$ 和 $\phi_w(k)$ 的更新公式如式(20)和(21)所示。

$$\mu_{w,d}(k)\propto[\sum_w x_{w,d}\mu_{w,d}(k)(1+B_{d,k})+\alpha-1]\times\frac{[\sum_d\sum_{w\in A_{w,w^*}}x_{w,d}\mu_{w,d}(k)+\beta-1]}{\sum_w[\sum_d\sum_{w\in A_{w,w^*}}x_{w,d}\mu_{w,d}(k)+\beta-1]} \tag{19}$$

$$\theta_d(k)=\sum_w x_{w,d}\mu_{w,d}(k)(1+B_{d,k}) \tag{20}$$

$$\phi_w(k)=\sum_d\sum_{w\in A_{w,w^*}} x_{w,d}\mu_{w,d}(k) \tag{21}$$

结合 SDS_TM 的图模型和更新公式，其训练过程如下所示，当模型初步收敛时(迭代次数大于下界 *bound*)，将 LDA 模型获得的结果与 word2vec 的单词局部语义分布结合，在主题建模过程中使用 GPU 模型进行语义强化。由于单词-单词之间相似度的计算时间较长，所以在模型初步收敛后，对矩阵 A_{w,w^*} 每次间隔一定次数(*iterval*)进行一次更新。其中向量 $\boldsymbol{v}$ 表示 word2vec 产生的单词局部语义分布表示，本文设置初步收敛下界 *bound*=30，更新间隔 *iterval*=50。

SDS_TM 的训练过程

输入： $x,K,T,\alpha,\beta,\boldsymbol{v}$

输出： θ_d,ϕ_k

1 随机为每个单词 $x_{w,d}$ 分配主题，并初始化以及标准化 $\mu^t_{w,d}(k)$，初始化 $\theta_d(k)$ 和 $\phi_w(k)$

2 for t=1 to T： //迭代循环，T 为循环次数

3 $\theta^t_d(k)\leftarrow 0$， $\phi^t_w(k)\leftarrow 0$ //对概率计数进行初始化

4 for $x_{w,d}$ in x : //遍历语料库中的每个单词

5 for k in K: //分别对每个主题进行更新

6 if $t<bound$:

7 使用式(12)更新 $\mu^t_{w,d}(k)$，使用式(13)和(14)更新 $\theta^t_d(k)$ 和 $\phi^t_w(k)$

8 else if $t>bound$:

9 if $t\%iterval==0$:

10 使用式(16)和(18)计算单词-单词和文档-单词的语义分布相似度矩阵

11 else:

12 使用式(19)更新 $\mu^t_{w,d}(k)$，使用式(20)和(21)更新 $\theta^t_d(k)$ 和 $\phi^t_w(k)$

13 $\theta_d(k)\leftarrow\theta^t_d(k)$，$\phi_w(k)\leftarrow\phi^t_w(k)$

14 //更新概率分布 $\theta_d(k)$ 和 $\phi_w(k)$

$$\theta_d(k)\leftarrow\frac{\theta_d(k)+\alpha-1}{\sum_k\theta_d(k)+\alpha-1},\phi_w(k)\leftarrow\frac{\phi_w(k)+\beta-1}{\sum_w\phi_w(k)+\beta-1}$$

3 实验分析

3.1 实验环境和数据集

本实验是在单机多核服务器上进行的，该服务器由 2 个 Intel(R)Xeon(R) CPU @2.10GHz 的 CPU 组成，每个 CPU 有 8 个核，总计 16 核，140GB 内存。

本文实验是在四个公开数据集上进行，分别为 Cora、WebKB、Reuters R8(R8)和 20 Newsgroups(20 News)数据集，文献[18]中对其进行了相关的介绍，表 2 展示了这四个数据集的相关信息描述。

表 2 数据集

数据集	*D*	*W*	*NNZ*	*Category*
Cora	2410	2961	103699	7
WebKB	4168	7764	202995	4
R8	7674	22931	322973	8
20 News	18821	92800	1549945	20

表 2 简要概括了这四个数据集，其中 *D* 为语料库中文档数、*W* 为单词表长度、*NNZ* 为非零元素个数，*Category* 为数据集中文本类别的数目。在实验之前，对数据集进行了一些预处理工作，主要包括去除标准的停用词、去除出现次数小于 3 的单词和词干化单词等。

在主题模型的研究和应用中，先验参数的选取对主题的建模产生一定的影响[19]，但是对于参数的研究不是本文的重点，所以为了保证对比实验的公平性和简单化，参考文献[1]中的参数设置，将所有算法中的先验参数都设置为 $\alpha=50/K$，$\beta=0.01$，其中 K 为主题个数，实验中总迭代次数设置为 T=1000，本文根据语义分布相似度来设置相应的相似度阈值，截取前 20%， $\lambda=0.6$，$\rho=0.4$，word2vec 模型的滑动窗口大小设置为 5。

3.2 评价标准

本文对主题模型的建模能力进行了评估，采用主题模型通用领域的性能评价指标：点互信息指数(Pointwise Mutual Information，PMI)[18,20]、分类准确率(Accuracy)[1]和混淆度(Perplexity) [1,7,21]。

点互信息(PMI)是衡量主题语义连贯性的常用评价指标，其主要思想是主题单词分布中概率值最高的前 *N* 个词更倾向于出现在语料库中的同一篇文档，PMI 评价指标通常与人工评价的结果一致，将主题中概率值最高的 *N* 个词之间的相关性作为 PMI 值，越高的 PMI 表示越强的主题语义连贯性，主题 *k* 的 PMI 计算公式如下所示：

$$PMI(k,W^k)=\frac{2}{N(N-2)}\sum_{i=2}^{N}\sum_{j=1}^{i-1}log\frac{Q(w_i^k,w_j^k)+\varepsilon}{Q(w_i^k)Q(w_j^k)} \tag{22}$$

其中， $Q(w)$ 表示单词 w 出现在语料库中的文档数目，$Q(w_i^k,w_j^k)$ 表示包含单词 $\{w_i,w_j\}$ 的文档数目， $W^k=(w_1^k,...,w_N^k)$ 为主题 k 中概率最大的 N 个单词列表，ε 是用来避免对数为 0 的一个小的正整数，本文设置 $N=10$， $\varepsilon=1$。

分类准确率是衡量文档语义表征能力的常用指标，将主题作为文档特征来实现文本分类，本文将数据集按 6:4 的比例划分为训练集和测试集，使用支持向量机(SVM)分类器实现分类任务，分别进行十次实验求其平均值作为准确率，不失一般性，经过实验验证，其他分类器的分类结果与其一致。分类准确率的计算公式为：

$$Acuracy=\frac{1}{|C|}\sum_{i\in C}\frac{T_i}{D_i} \tag{23}$$

其中，$|C|$ 表示文本类别的数目，D_i 表示类别 i 中的文本数目，T_i 表示类别 i 中被分类正确的文本数目。

混淆度是评价 LDA 模型建模好坏的常用评价指标之一，其可以被理解为语料库中所有单词似然值几何平均数的倒数，越低的混淆度表示越好的泛化性能。其计算公式为：

$$Perp = \exp\left\{-\frac{\sum_{w,d} x_{w,d} log[\sum_k \theta_d(k)\phi_w(k)]}{\sum_{w,d} x_{w,d}}\right\} \quad (24)$$

3.3 实验对比分析

3.3.1 语义连贯性分析

本文将目前流行的 LDA 推理算法，即变分贝叶斯(VB)、吉布斯采样(GS)和期望最大化(EM)与提出的基于语义分布相似度的主题模型(SDS_TM)作对比，图 4 展示了四种算法在不同的主题数 K 下，Cora、WebKB、R8 和 20 News 数据集上的 PMI 值对比，可以看出，本文提出的 SDS_TM 的 PMI 值总体较高，表明其抽取的主题具有较高的语义连贯性。

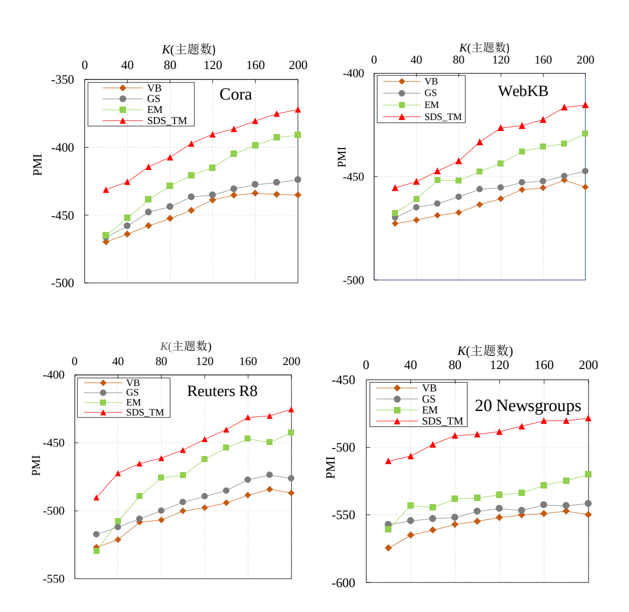


图 4 不同主题数 K 下 PMI 值比较

VB 算法优化的是后验概率的近似下界，GS 算法是通过简单采样方法获得单词的主题标签，这两种都是近似推断，EM 算法是精确地优化后验概率表示，所以较 VB 和 GS 算法能够获得语义相关性更强的主题，但是这三种推理算法都是建立在词袋(BOW)模型的基础上，忽略了主题模型中的语义关系，而 SDS_TM 能够有效地将单词-单词和文档-主题的语义关联融合到主题建模中，因此 SDS_TM 能够获得语义连贯性较高的主题。

3.3.2 文本分类效果分析

本文将 SDS_TM 模型用于文本分类任务，以验证模型整体有效性，文本分类准确率越高，则表示主题的特征表达能力越强。表 3 展示了四种算法在 R8 数据和 20 News 数据集上文本分类准确率随着主题数 K 的变化情况， 可以看出 SDS_TM 模型在两个数据集上都能获得较高的准确率，其中精确推理算法 EM 算法比近似推断算法 VB 和 GS 算法的准确率较高，其中 R8 数据集上分类准确率较 20 News 数据集上较高，这可能由于文档大小对主题建模的影响，R8 比 20 News 具有较短的词汇表，在 R8 数据上文本的稀疏性较小，能够获取更加相似语义信息，更能有效地引导主题建模。

表 3 不同主题数 K 下分类准确率比较

数据集	算法	K=20	K=40	K=60	K=80	K=100
R8	VB	0.784	0.778	0.772	0.774	0.766
	GS	0.905	0.905	0.881	0.886	0.883
	EM	0.909	0.901	0.888	0.893	0.887
	SDS_TM	**0.937**	**0.921**	**0.919**	**0.924**	**0.936**
20 News	VB	0.526	0.557	0.557	0.5432	0.5458
	GS	0.598	0.720	0.718	0.729	0.724
	EM	0.710	0.754	0.757	0.737	0.733
	SDS_TM	**0.735**	**0.781**	**0.781**	**0.753**	**0.778**

3.3.3 算法收敛性和模型精度分析

收敛性是一种评价模型训练快慢的常用指标，图 5 和图 6 展示了四种 LDA 算法在数据集 R8 和 20 News 上混淆度随迭代次数的变化情况。由于 SDS_TM 模型在建模过程中有效地融合了语义分布相似度信息，所以其收敛速度最快，其中，VB 算法和 EM 算法为每个单词保留了所有的主题信息，它们的收敛速度较 GS 算法较快，GS 算法只为每个单词采样出一个主题，且采样的过程比较慢，所以其收敛速度最慢。此外，SDS_TM 模型比其他三种算法在最终的混淆度方面存在优势，混淆度越低则模型精度越高，其在未知数据集上具有更强的泛化能力，VB 算法收敛后的混淆度最大，因为其简化了模型的复杂度，所以造成了模型精度的损失。在迭代次数大于 30 时，主题模型趋向于大致收敛，通过引入单词-单词和文档-主题的语义分布相似度来引导主题建模，SDS_TM 模型的混淆度下降幅度增加，并快速趋向于收敛状态，因此 SDS_TM 在收敛速度和模型精度方面较其他算法都能够表现得更加优越。

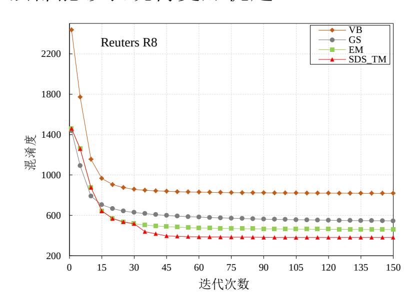


图 5 R8 数据集上混淆随迭代次数变化

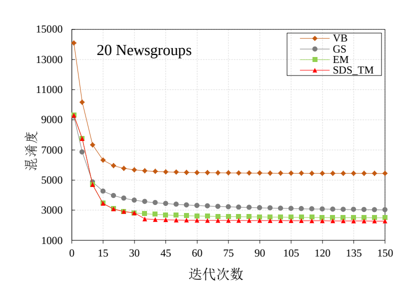


图 6 20 News 数据集上混淆随迭代次数变化

4 结束语

本文针对目前的主题模型推理算法中存在的语义连贯性较

差，文档表征能力不强等缺点，提出了一种基于语义分布相似度的主题模型(SDS_TM)。此模型在 EM 算法框架下，有效地使用 GPU 模型将单词-单词和文档-主题之间的语义关联融合到主题建模过程中，从而推断出主题模型的参数。实验表明，SDS_TM 在主题语义连贯性、文本分类准确率、收敛速度和模型精度方面均表现优异。下一步针对 SDS_TM 的研究主要集中在提高计算语义分布相似度的速度，及其在大数据流上的应用和并行加速等方面，在提高模型精度的情况下，加快模型的训练速度。

参考文献：

[1] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Joural of Machine Learning Research, 2003, 3: 993-1022.

[2] Hoffman M D, Blei D M, Bach F R. Online Learning for Latent Dirichlet Allocation [J]. Advances in Neural Information Processing Systems, 2010, 23: 856-864.

[3] Dunlavy D M, O'Leary D P, Conroy J M, *et al.* QCS: A system for querying, clustering and summarizing documents [J]. Information Processing & Management, 2007, 43 (6): 1588-1605.

[4] Niebles Juan Carlos, Wang Hongcheng, Li Feifei. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words [J]. International Journal of Computer Vision, 2008, 79 (3): 299-318.

[5] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proceedings of the National academy of Sciences, 2004, 101 (Suppl 1): 5228-5235.

[6] Liu Xiaosheng, Zeng Jia, Yang Xi, *et al.* Scalable Parallel EM Algorithms for Latent Dirichlet Allocation in Multi-Core Systems [C]// International Conference on World Wide Web. New York: ACM Press, 2015: 669-679.

[7] Zhang JianWei, Zeng Jia, Yuan Mingxuan, *et al.* LDA Revisited: Entropy, Prior and Convergence [C]// Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2016: 1763-1772.

[8] Foulds J, Boyles L, Dubois C, *et al.* Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation [C]// Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2013: 446-454.

[9] Rosen-Zvi M, Griffiths T, Steyvers M, *et al.* The author-topic model for authors and documents [C]// Proceedings of the 20th Conference on Uncertainty in Artifical Intelligence. 2012: 487-494.

[10] Chen Zhiyuan, Mukherjee Arjun, Liu Bing, *et al.* Discovering coherent topics using general knowledge [C]// Proceedings of the 22th ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2013: 209-218.

[11] Bekoulis G, Rousseau F. Graph-Based Term Weighting Scheme for Topic Modeling [C]// International Conference on Data Mining Workshops. New York: IEEE, 2016: 1039-1044.

[12] Mimno D, Wallach H M, Talley E, *et al.* Optimizing Semantic Coherence in Topic Models [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011: 262-272.

[13] Kotz S, Mahmoud H, Robert P. On generalized Polya urn models [J]. Statistics Probability Letters, 2000, 49 (2): 163-173.

[14] Gilks W R. Markov Chain Monte Carlo [M]. Numerical Analysis for Statisticians. New York: Springer, 1999: 238-245.

[15] Asuncion A, Welling M, Smyth P, *et al.* On smoothing and inference for topic models [C]// Conference on Uncertainty in Artificial Intelligence. 2009: 27-34.

[16] Andrzejewski David, Zhu Xiaojin, Craven Mark. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors [C]// International Conference on Machine Learning. New York: ACM Press, 2009: 25-32.

[17] Mikolov Tomas, Sutskever Ilya, Chen Kai, *et al.* Distributed Representations of Words and Phrases and their Compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[18] Wu Xiaona, Zeng Jia, Yan Jianfeng, *et al.* Finding Better Topics: Features, Priors and Constraints [C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. New York: Springer, 2014: 296-310.

[19] Wallach H M, Mimno D M, Mccallum A. Rethinking LDA: Why priors matter [C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2009: 1973-1981.

[20] Newman D, Lau J H, Grieser K, *et al.* Automatic evaluation of topic coherence [C]// Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010: 100-108.

[21] 常东亚，严建峰，杨璐. 基于中心词的上下文主题模型 [J]. 计算机应用研究, 2018, 35 (4): 1005-1009. (Chang Dongya, Yan Jianfeng, Yang Lu. Centroid-word based context topic model [J]. Application Research of Computers, 2018, 35 (4): 1005-1009.)